Methods to Estimate Medical Claims Data Using Regression

Scarlett Townsend

California State University, East Bay

Abstract

A dataset containing seven variables, three of them categorical and 1,338 observations was utilized to explore a variety of regression techniques and the difficulties implicit in applying regression to medical claims data. Due to excessive violations of the normality of residuals and non-constant variance, segmentation guided by visualization and response variable transformation was used to help fit a model. Additional thoughts regarding practical implementation and methods to supplement the analysis are also considered.

Methods to Estimate Medical Claims Data Using Regression

Introduction

Anticipation of claims data is vital to many businesses but especially important for insurers who must price policies guaranteeing payment for these claims. For this analysis, data was leveraged from Packt Publishing consisting of hypothetical medical expenses for patients in the United States. The data was created using demographic statistics from the U.S. Census Bureau to reflect actual conditions (Lantz). The dataset includes seven variables, three of them categorical and 1,338 observations.

Materials and Methods

As indicated, the target variable is annual medical expenses. Looking at the data provided, the expenses range from $1,122 to $63,770. The average is just over $13,000 and the median is $9,382 indicating there are some outliers pulling the average up. The standard deviation of expenses is around 12,000 and in reviewing the plots, significant skewness is apparent. While the majority of expenses are between $0 and $15,000, fully 10% of the expenses data points are high values, in excess of $34,000.

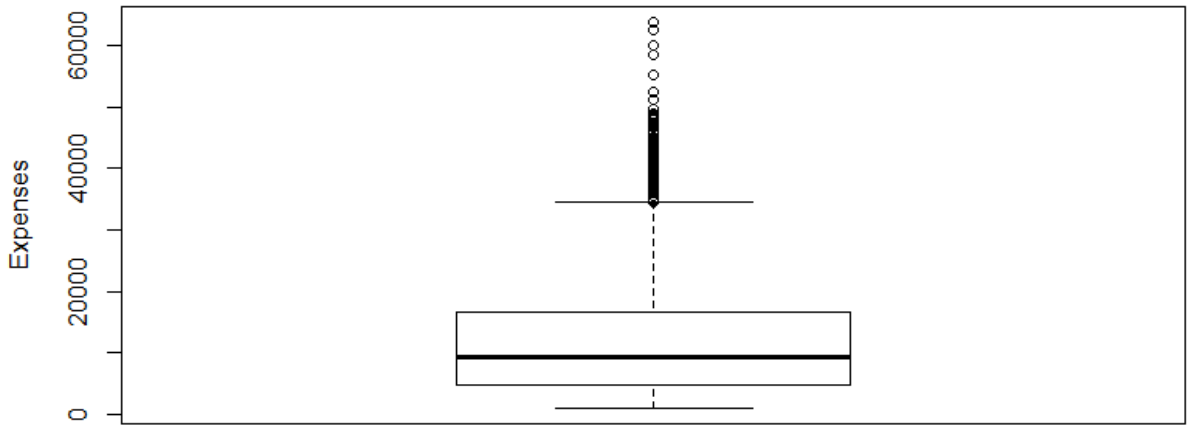

Figure 1. Boxplot of Expenses.

Looking at the predictor variables, Age of the insured is included and ages range from 18 to 64 with a standard deviation of 14 and median of 39, thus spanning the ages of working age adults. From the histogram (Figure 1), the distribution appears somewhat uniform with slightly more instances below 18 and fewer over 60. The data shows only discrete or rounded ages.

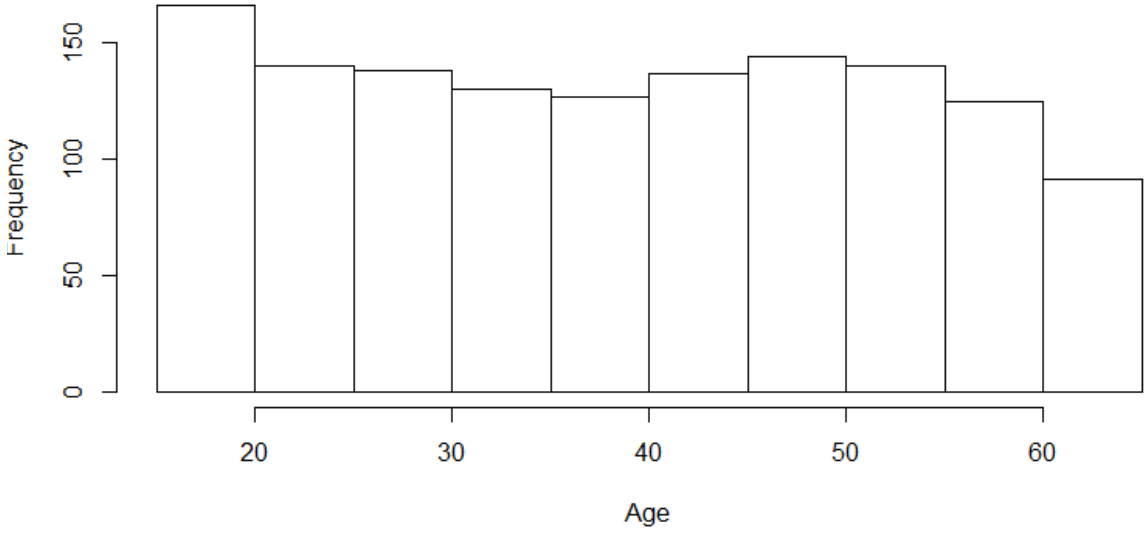

Figure 2. Histogram of Age.

BMI has a more normal distribution, slightly skewed to the right. The range is from 16 to 53.10 with a median at 30.4 and standard deviation of 6. The data set shows 9 values with extreme high BMIs over 47.4.

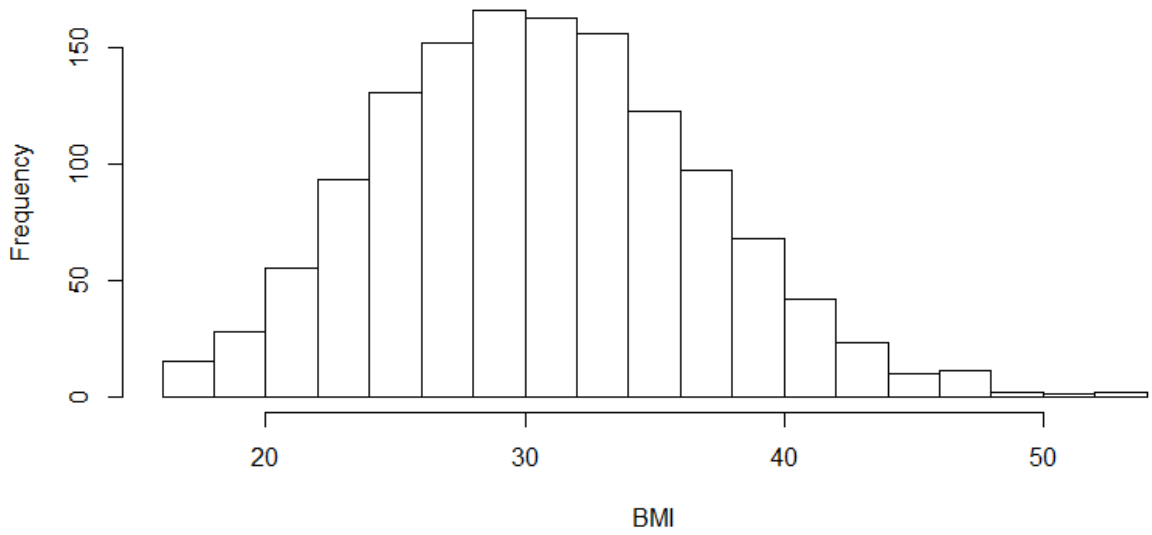

Figure 3. Histogram of BMI.

Another discrete datapoint, the number of children is also considered as a potential predictor of medical expenses. We see the range is between 0 and 5 with a median of 1.

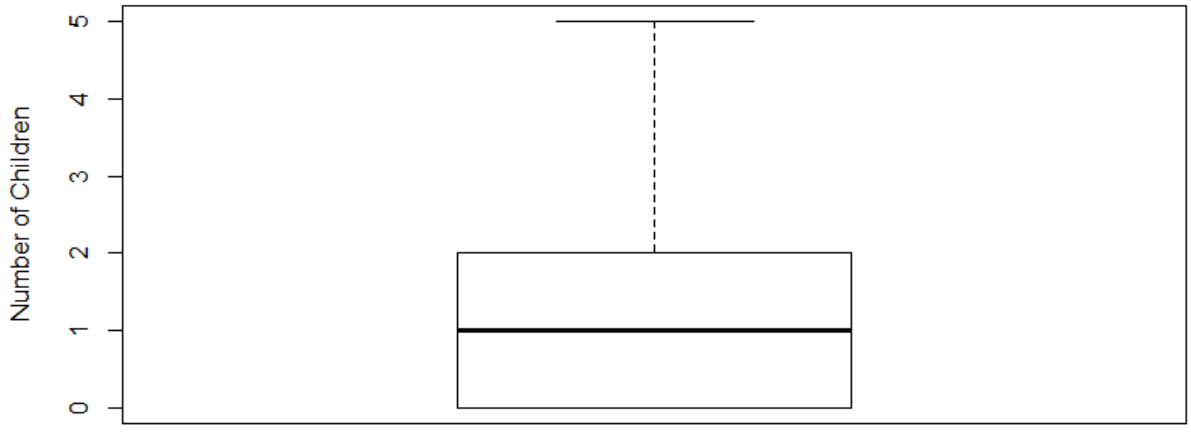

Figure 4. Boxplot of Number of Children.

Categorical predictor variables considered are sex, smoking status and region. Sex is represented with a nearly even split; there are 662 females and 676 males. Approximately 20%

of the observations are classified as smokers. Regionally there are 324 from the northeast, 325 from the northwest, 364 from the southeast and 325 from the southwest.

Examining the correlations between variable, it is noted that age has the strongest linear relationship with expenses with a correlation coefficient of 0.3, followed by BMI at 0.2. The number of children does not have a strong linear relationship with expenses with a coefficient of 0.07. Between the predictor variables, age and BMI are most correlated with a coefficient of 0.1.

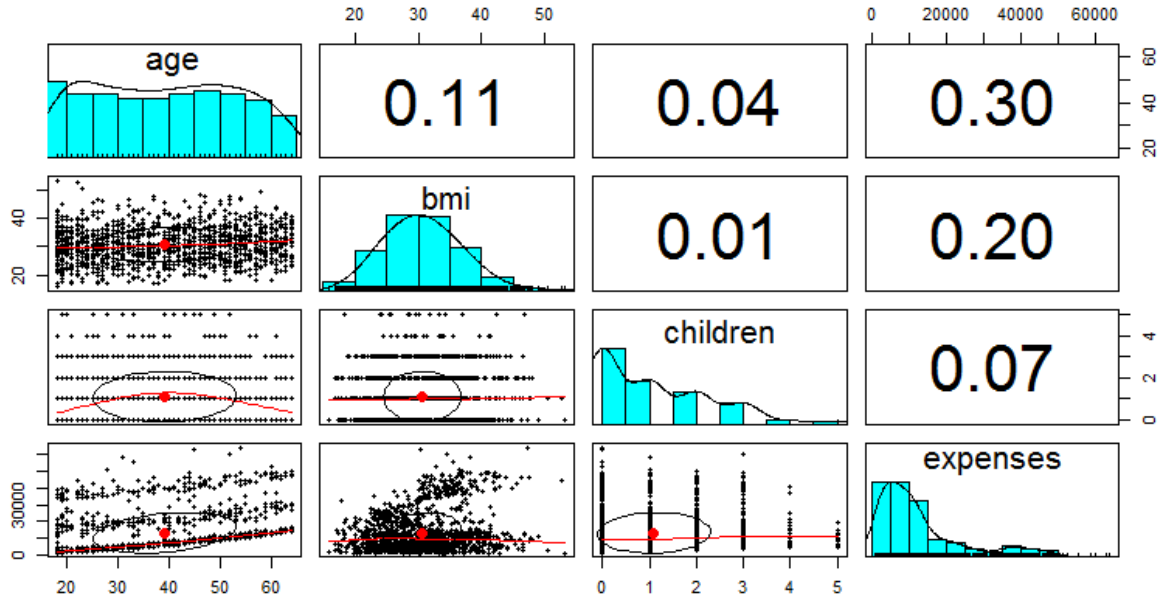

Figure 5. Scatterplot Matrix

Examining plots color coded with the categorical data, a distinct pattern emerges. In the BMI plot, the linear relationship between BMI and expenses appears to be somewhat consolidated for nonsmokers whereas for smokers it segments into two groups with differing intercepts based on a BMI threshold of around 30.

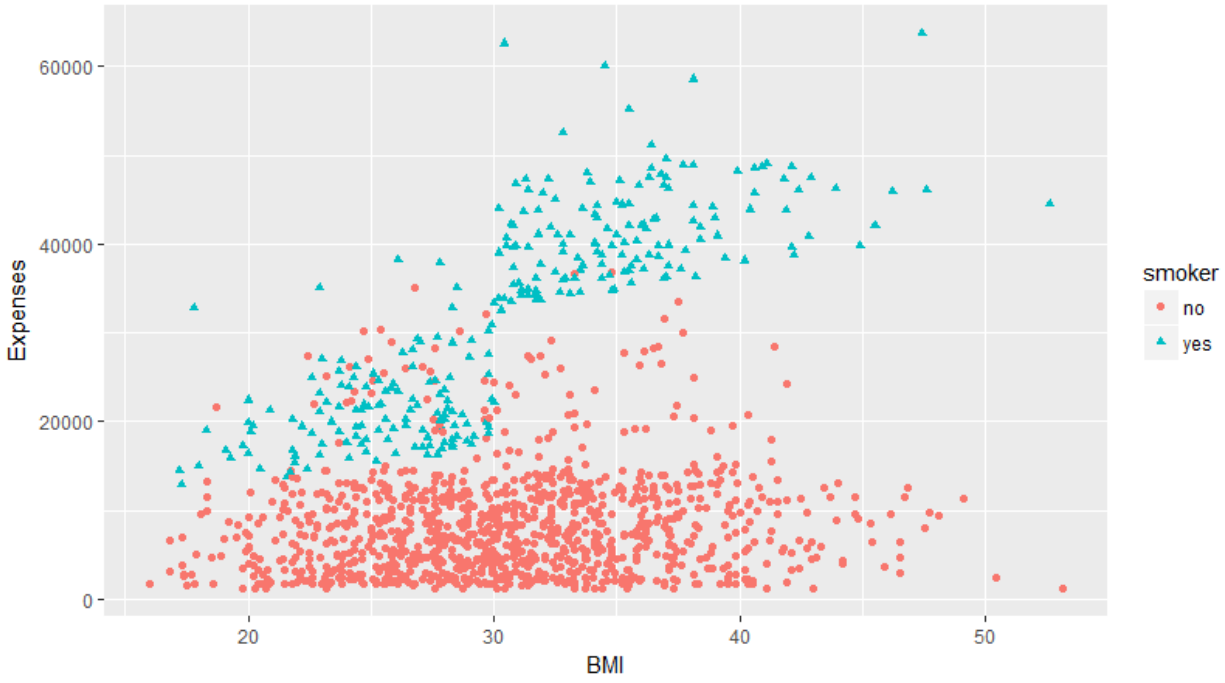

Figure 6. Scatterplot of BMI and Expenses by Smoking Status

Creating a group for smokers with BMI above 30 and another for those with lower BMI, the trend becomes more evident.

Figure 7. Scatterplot of Age and Expenses by Smoking and BMI group

Results

Single-factor and Untransformed Models

As an initial exploration an untransformed additive model and single-variable linear models were fitted. The AIC for the single-variable models, shown in Figure 8 confirms that the group variable created results in a model with a lower AIC, thus a better fit than any other predictor considered alone. Figure 9 shows the Box-Cox analysis indicating a maximum likelihood of a power for transforming the response variable of less than 0.5, which suggests a logarithmic transformation (Kutner, Nachtsheim & Neter).

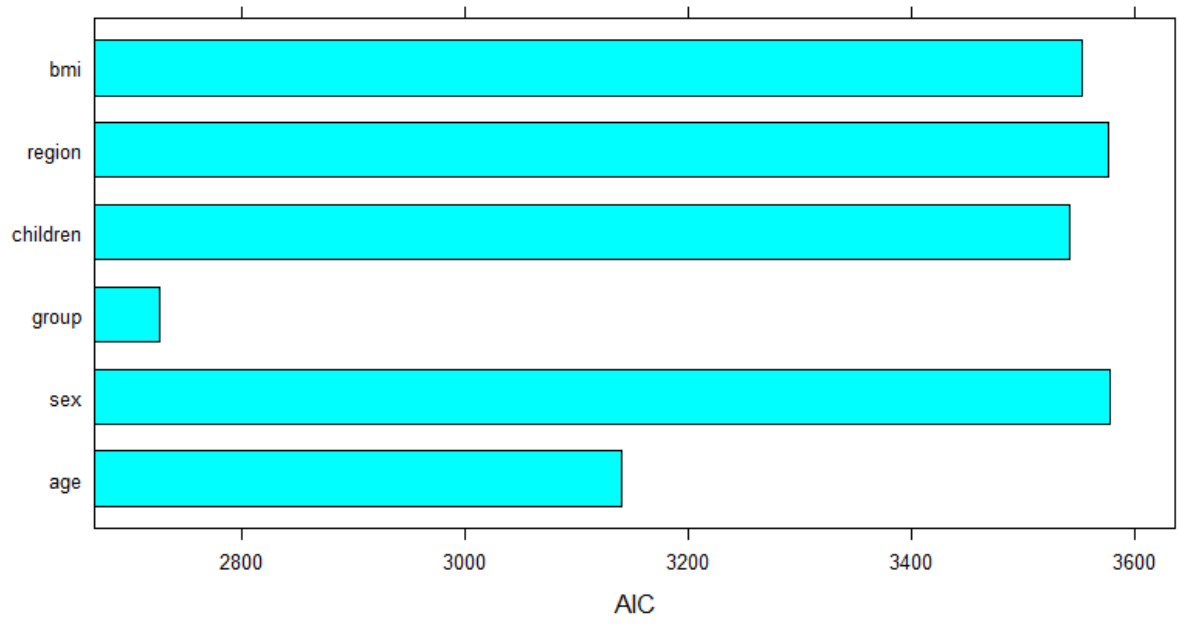

Figure 8. AIC of Single-factor Models

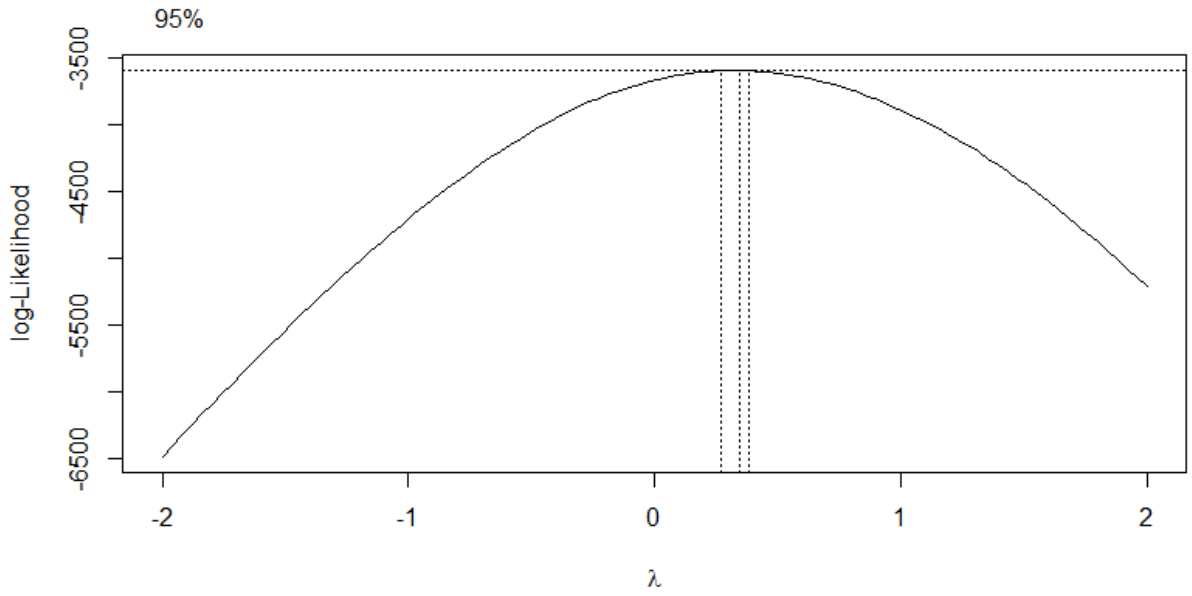

Figure 9. Box-Cox Analysis of the Additive Model

Transformed Model

To improve the fit, a logarithmic transformation was selected to be applied to the response variable. The ANOVA in *Table* 1 shows all predictor variables are significant and contribute to the ability to predict expenses. Note that smoker was omitted because it was incorporated into the group variable. Based on this model, males are anticipated to have less expenses which seems to correspond to the expectation. Other interesting results include, the northeast is expected to have the highest claims and an increase in the number of children, increases the claims. The expected relationship between smoking and high BMI increasing medical costs is also evident. The adjusted R squared indicates 79% of the variation in expenses is explained by this model.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.2967016  0.0733119  99.530  < 2e-16 ***
age                0.0347481  0.0008347  41.629  < 2e-16 ***
sexmale           -0.0871992  0.0233762  -3.730 0.000199 ***
```

```
groupSmoker High BMI  1.8507293  0.0394124  46.958  < 2e-16 ***
groupSmoker Low BMI   1.2330181  0.0409607  30.102  < 2e-16 ***
regionnorthwest      -0.0593308  0.0334074  -1.776 0.075966 .
regionsoutheast      -0.1494601  0.0335819  -4.451 9.28e-06 ***
regionsouthwest      -0.1349861  0.0335253  -4.026 5.98e-05 ***
children              0.1025558  0.0096655  10.611  < 2e-16 ***
bmi                   0.0045880  0.0021562   2.128 0.033537 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4252 on 1328 degrees of freedom
Multiple R-squared:  0.7876,    Adjusted R-squared:  0.7862
F-statistic: 547.3 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Table 1. ANOVA for Transformed Model

The transformed model's residual plots show a curved pattern. In addition, the Q-Q plot indicated significant normality violations. These visual observations of assumption violations were confirmed with the formal Levene test of equal variances as well as the Shapiro-Wilks test of normality of the residuals.

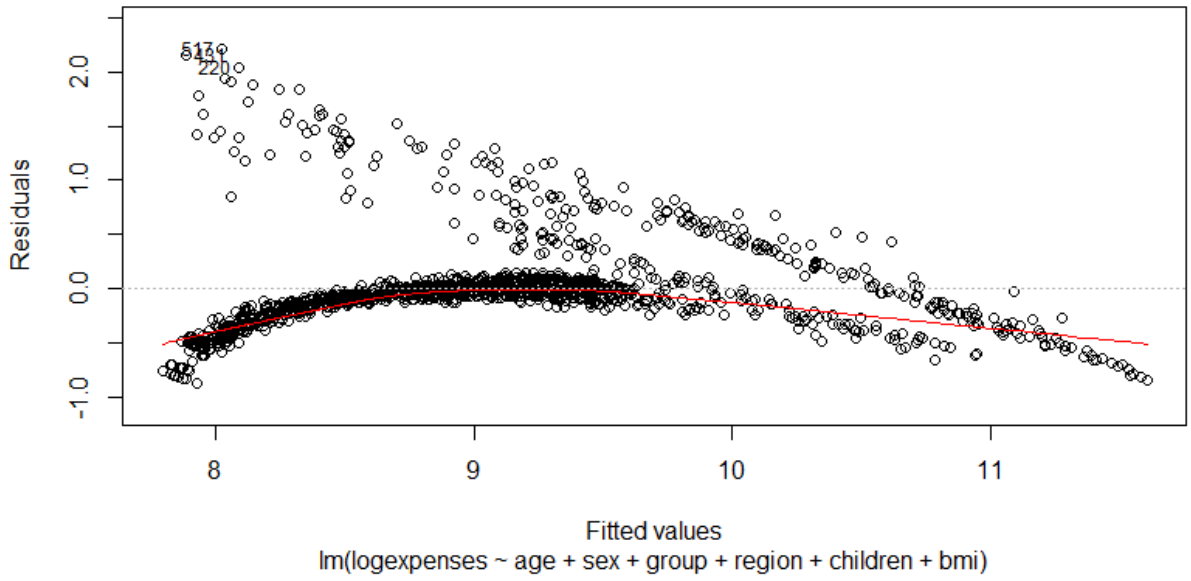

Figure 8. Scaterplot of Fitted Values vs Residuals for Transformed Model.

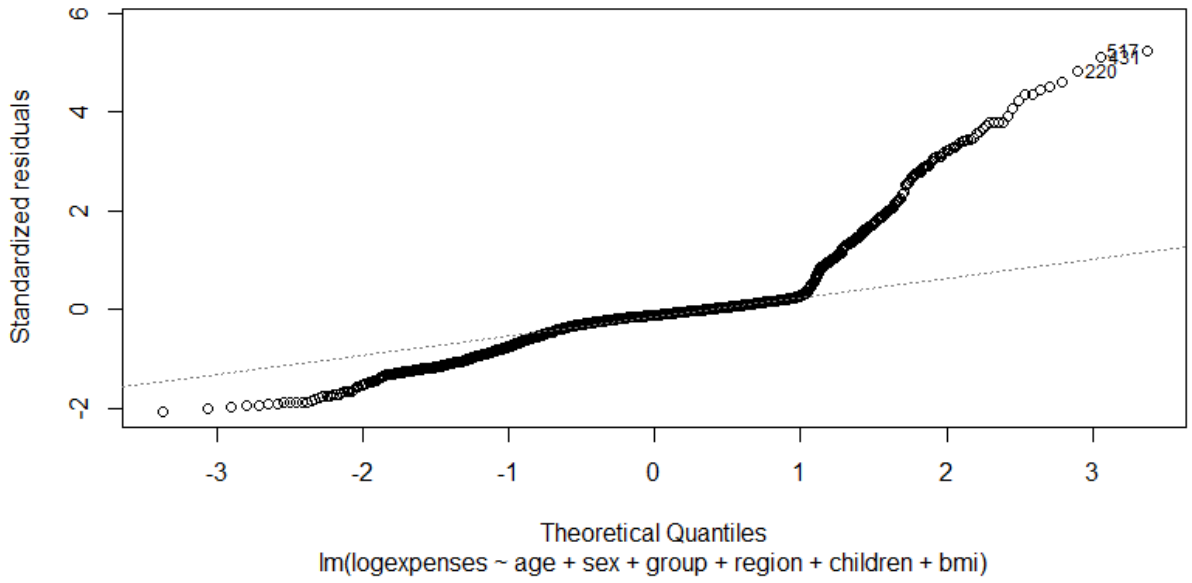

Figure 9. Quantile-Quantile Plot for Transformed Model

Because of the nonlinear pattern, a second order age term was included. The ANOVA in Table 2 shows an increase in the adjusted R squared to 79%. Unfortunately the regression assumptions of normality of residuals was not supported in this case either.

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.970e+00  1.208e-01  57.702  < 2e-16 ***
age                   5.388e-02  5.689e-03   9.471  < 2e-16 ***
age2                 -2.413e-04  7.096e-05  -3.400 0.000695 ***
sexmale              -8.681e-02  2.328e-02  -3.728 0.000201 ***
groupSmoker High BMI  1.851e+00  3.926e-02  47.157  < 2e-16 ***
groupSmoker Low BMI   1.231e+00  4.080e-02  30.169  < 2e-16 ***
regionnorthwest      -5.840e-02  3.328e-02  -1.755 0.079490 .
regionsoutheast      -1.496e-01  3.345e-02  -4.474 8.34e-06 ***
regionsouthwest      -1.352e-01  3.339e-02  -4.048 5.47e-05 ***
children              9.233e-02  1.009e-02   9.154  < 2e-16 ***
bmi                   4.798e-03  2.149e-03   2.233 0.025717 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4235 on 1327 degrees of freedom
Multiple R-squared:  0.7895,   Adjusted R-squared:  0.7879
F-statistic: 497.6 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Other Fits: Comparing AIC, BIC and Adjusted R Squared

Multiple interaction terms were explored but none seemed to improve the fit or the normality of the residuals significantly, other than the interaction captured by the group variable.

Beginning with the variable producing the lowest AIC of all single-factor models and sequentially adding the other terms, we can see how adding additional factors reduces AIC and increases the adjusted R-squared. There was no discrepancy in this case between relative AIC and relative BIC. The untransformed model is indicated with the red lines in Figures 10 and 12, interestingly, the adjusted R-squared is higher for the untransformed model.

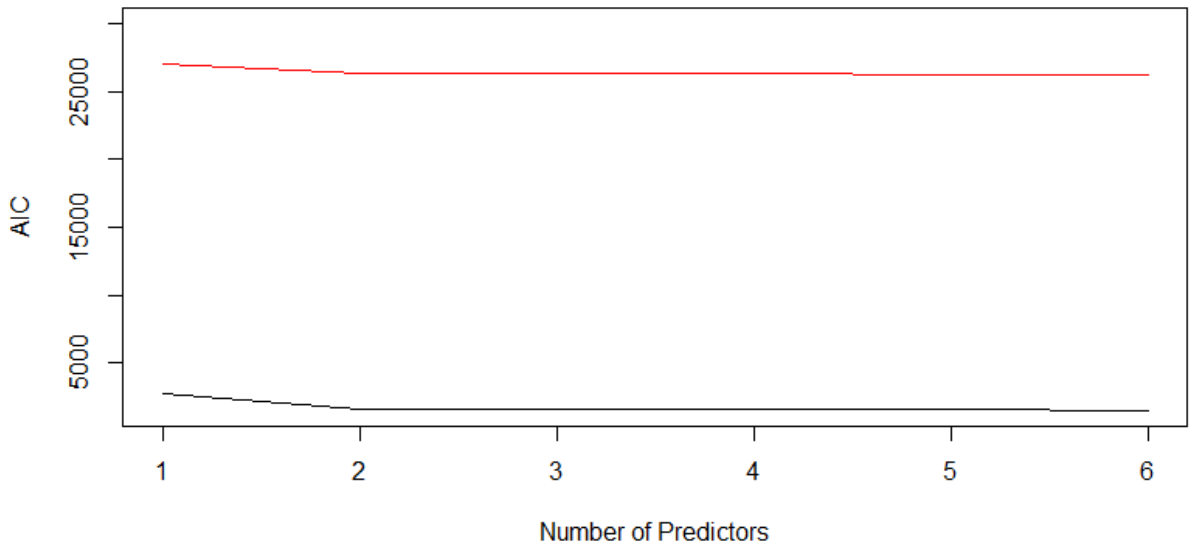

Figure 10. AIC by Number of Predictors (Untransformed Model in Red)

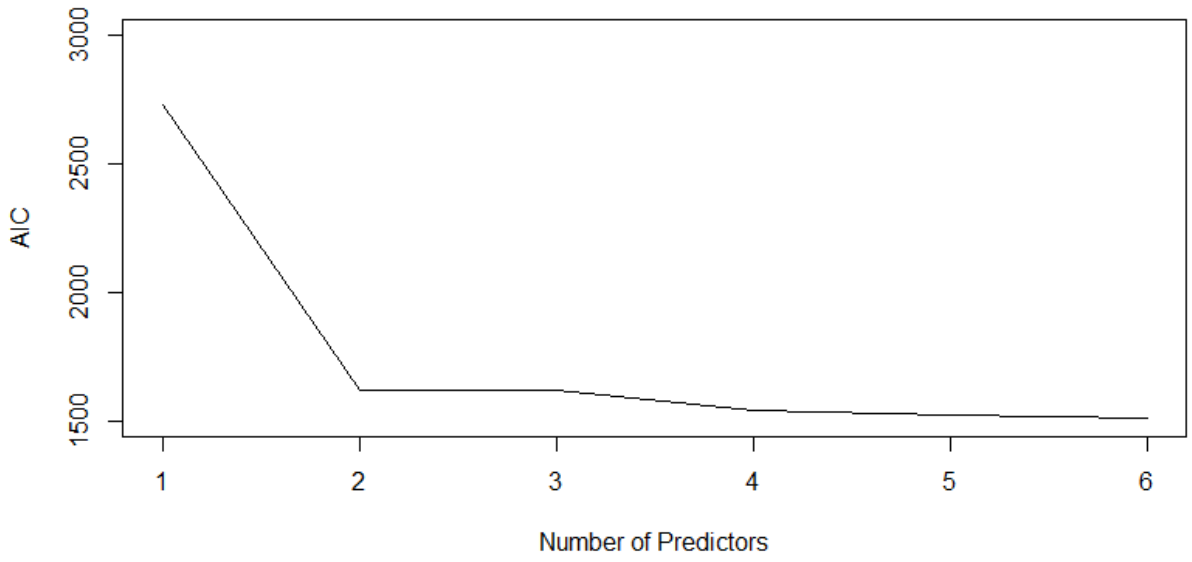

Figure 11. AIC for Logarithmic Transformation with Second-Order Age Included by Number of Predictor Variables

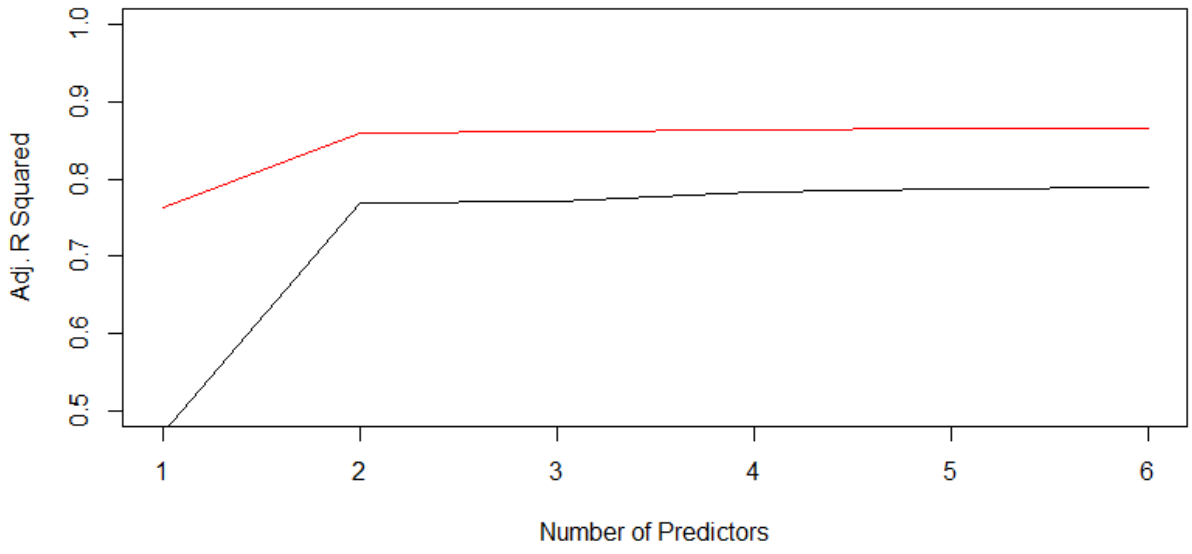

Figure 12. Adjusted R-Squared by Number of Predictors (Untransformed Model in Red)

Discussions

Although these models cannot be used reliably because they have been shown to violate the assumptions underlying regression, often it is the case in real-world situations, the data cannot be coerced into textbook regularity. There are nonparametric option that do not rely on normality to fit the data that can be considered in place of these models or alongside them when predicting the costs of medical expenses including lowess and regression trees (Kutner, Nachtsheim & Neter). These techniques offer a localized smoothing of the curve but would not necessarily address the problem of excessive outliers such as seen in this data set.

The visualization shows clear linear patterns with many outliers, a strong indication that there is another, unknown predictor variable influencing the expenses. Considering this, it would be recommended that the modelers make attempts to collect additional data from the subjects. If that could not be done and a regression was strongly preferred, one solution would be to group the outliers into another model and then determine the frequency of occurrence of the deviations from the main pattern. The statistician could give results such as, 80% of customers are expected to fall into this major regression pattern and thus be predicted by a formula derived without the outliers whereas the other 20% are expected to fall into the alternate pattern. This would allow for costs to be weighted to arrive at a way to reasonably set pricing or budgets.

References

Kutner, M., Natchtsheim C., & Neter J. (2004). Applied Linear Regression Models.

Lantz, B. (2015). Machine Learning with R.

Prof - Deep brought this dataset to us and mentioned it was from his statistical learning class so, it seemed reasonable. When I was looking for the reference to include I found the book which walks through regression in R on the same dataset.

I did my analysis prior to knowing about this and you can see there are many differences in the approach and it's not nearly as robust as what we learned in your class. I would not have selected this data if I had been aware.

Appendix: R Code

```
library(readr)
ins <- read_csv("C:/Users/Scarlett/Google Drive/201501_CSUEB/201704_Spring_Regression/insurance.csv")
#data descriptions
ins$sex = as.factor(ins$sex)
ins$group = as.factor(ins$group)
ins$region = as.factor(ins$region)
summary(ins)

hist(ins$expenses, main = "",xlab="Expenses")
expensesBox=boxplot(ins$expenses, ylab="Expenses")
length(expensesBox$out)/length(ins$expenses)
min(expensesBox$out)
sd(ins$expenses)

hist(ins$age, main = "",xlab="Age",breaks=(64-18)/5)
sd(ins$age)
sd(ins$age)/(max(ins$age)-min(ins$age))

hist(ins$bmi, main = "",xlab="BMI",breaks=20)
bmiBox=boxplot(ins$bmi)
min(bmiBox$out)
sd(ins$bmi)
sd(ins$bmi)/(max(ins$bmi)-min(ins$bmi))

hist(ins$children, main = "",xlab="Children",breaks=5)
min(childrenBox$out)
sd(ins$children)
```

```
sd(ins$children)/(max(ins$children)-min(ins$children))


length(subset(ins$group,ins$group=='yes'))/length(ins$group)
plot(ins$region, xlab="Region")


#data analysis
cor(ins[,c(1,3,4,7)])
pairs(ins)
install.packages("psych")
library(psych)
pairs.panels(ins[,c(1,3,4,7)])
library(ggplot2)
qplot(bmi,expenses, colour = group, shape = group, data = ins, ylab="Expenses",
xlab="BMI")
ins$group=ifelse(ins$group=='yes'& ins$bmi>30,'group High BMI',
            ifelse(ins$group=='yes','group Low BMI','Nongroup'))
qplot(age, expenses, colour = group, shape = group, data = ins)


#a - untransformed, all variables, no interactions
a=lm(expenses~age+sex+group+region+children+bmi, data=ins)
plot(a)
shapiro.test(a$residuals)
boxcox(a)
# see about dropping
library(MASS)
aa <- stepAIC(a, direction="both")
aaa=aa$anova # display results
aaa
plot(aaa$AIC,xlab="Step",ylab="AIC",type="l")
```

```
#determine transformation

boxcox(a)$x[which(boxcox(a)$y==max(boxcox(a)$y))]# 0.1010101

boxcox(lm(expenses ~ age, data = ins))$x[which(boxcox(lm(expenses ~ age, data =
ins))$y

        ==max(boxcox(lm(expenses ~ age, data = ins))$y))]#-0.1010101

boxcox(lm(expenses ~ bmi, data = ins))$x[which(boxcox(lm(expenses ~ bmi, data =
ins))$y

        ==max(boxcox(lm(expenses ~ bmi, data = ins))$y))]#0.06060606

boxcox(lm(expenses ~ children, data = ins))$x[which(boxcox(lm(expenses ~ children,
data = ins))$y

        ==max(boxcox(lm(expenses ~ children, data = ins))$y))]#0.02020202

#transform

ins$logexpenses = log(ins$expenses, base=exp(1))

ins$age2= ins$age*ins$age

tr = lm(logexpenses ~ age+sex+group+region+children+bmi, data = ins)

summary(tr)

plot(tr)

tr2 = lm(logexpenses ~ age+age2+sex+group+region+children+bmi, data = ins)

summary(tr2)

plot(tr2)

#for visuals

a2=lm(logexpenses~age, data=ins)

a3=lm(logexpenses~sex, data=ins)

a4=lm(logexpenses~group, data=ins)

a5=lm(logexpenses~children, data=ins)

a6=lm(logexpenses~region, data=ins)

a7=lm(logexpenses~bmi, data=ins)


y=c(AIC(a2),AIC(a3),AIC(a4),AIC(a5),AIC(a6),AIC(a7))

names(y)=c('age','sex','group','children','region','bmi')

library(lattice)
```

```
barchart(y, xlab='AIC')

a=lm(expenses~group, data=ins)
b=lm(expenses~group+age+age2, data=ins)
c=lm(expenses~group+age+age2+bmi, data=ins)
d=lm(expenses~group+age+age2+bmi+children, data=ins)
e=lm(expenses~group+age+age2+bmi+children+region, data=ins)
f=lm(expenses~group+age+age2+bmi+children+region+sex, data=ins)

aa=lm(logexpenses~group, data=ins)
bb=lm(logexpenses~group+age+age2, data=ins)
cc=lm(logexpenses~group+age+age2+bmi, data=ins)
dd=lm(logexpenses~group+age+age2+bmi+children, data=ins)
ee=lm(logexpenses~group+age+age2+bmi+children+region, data=ins)
ff=lm(logexpenses~group+age+age2+bmi+children+region+sex, data=ins)


y=c(AIC(a),AIC(b),AIC(c),AIC(d),AIC(e),AIC(f))
y2=c(BIC(a),BIC(b),BIC(c),BIC(d),BIC(e),BIC(f))
y3=c(summary(a)$r.squared,summary(b)$r.squared
    ,summary(c)$r.squared,summary(d)$r.squared
    ,summary(e)$r.squared,summary(f)$r.squared)

z=c(AIC(aa),AIC(bb),AIC(cc),AIC(dd),AIC(ee),AIC(ff))
z2=c(BIC(aa),BIC(bb),BIC(cc),BIC(dd),BIC(ee),BIC(ff))
z3=c(summary(aa)$r.squared,summary(bb)$r.squared
    ,summary(cc)$r.squared,summary(dd)$r.squared
    ,summary(ee)$r.squared,summary(ff)$r.squared)

plot(z,type='l',ylab="AIC",xlab='Number of Predictors',ylim=c(1500,30000))
```

```
lines(y,col='red')
plot(z,type='l',ylab="AIC",xlab='Number of Predictors',ylim=c(1500,3000))
plot(z3,type='l',ylab='Adj. R Squared',xlab='Number of Predictors',ylim=c(.5,1))
lines(y3,col="red")
```